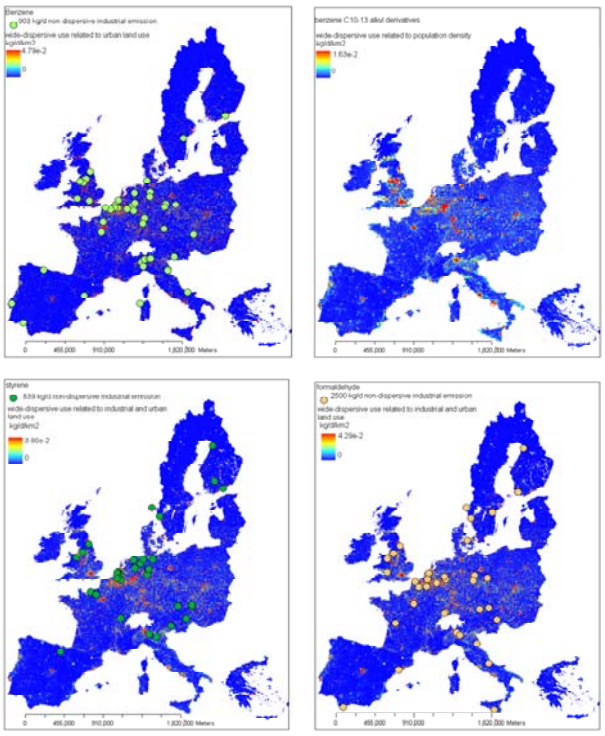
JRC Scientific and Technical Reports

Using decision tree analysis and GIS in Modelling (semi)VOC Emissions at the European Scale

Patrik Fauser, Alberto Pistocchi, Marianne Thomsen



EUROPEAN COMMISSION

Institute for Environment and Sustainability

EUR 24254 EN - 2010

The mission of the JRC-IES is to provide scientific-technical support to the European Union's policies for the protection and sustainable development of the European and global environment.

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information
Address: Alberto Pistocchi, JRC, TP 460, Via Fermi 2749, 21027 Ispra (VA), Italy
E-mail: alberto.pistocchi@jrc.ec.europa.eu
Tel.: 00 39 0332 785591
Fax: 00 39 0332 785601

http://ies.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC 56445

EUR 24254 EN
ISBN 978-92-79-15023-4
ISSN 1018-5593
doi:10.2788/62874

Luxembourg: Publications Office of the European Union

Printed in Italy

1. Introduction

Chemicals in the environment arising from human emissions pose issues concerning human health and environmental risks (e.g. [1], [2]). Emission quantification is the natural starting point of the life-cycle analysis of chemicals and is key to any modeling effort aimed at predicting chemical concentration. Emissions are often related to production and use of chemicals and the complex nature of chemical emission patterns makes quantification of emissions very uncertain. In many cases the predominant uncertainties in a risk assessment are indeed related to the uncertainties of the emission inventories.

There are many ways to perform an emission inventory. The emission inventory guidebook prepared by the UNECE/EMEP Task Force on Emissions Inventories and Projections [3] to support reporting under the UNECE Convention on Long-Range Transboundary Air Pollution [4] and the EU directive on national emission ceilings 2001/81/EC provides a comprehensive state-of-the-art methodological guide for atmospheric emissions.

A comprehensive emission estimate has been done, for single chemicals, in the context of risk assessments made for existing chemicals by the EU member states and coordinated by the European Chemicals Bureau (ECB) before entry into force of the EC Regulation N. 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH); this work is accessible in the form of publicly available risk assessment reports (RARs) ([5]). Existing chemicals are defined in the European Union (EU) as any chemical substance listed in the European INventory of Existing Commercial Substances (EINECS), an inventory containing more than 100,000 substances ([6]). Each of these chemicals may pose a risk to humans and the environment, and may therefore be potentially subjected to risk assessment.

The ECB RARs provide information for individual chemicals and include data, modeling results and expert judgments, based on IUCLID ([7]), a tool for data collection and evaluation within the EU-Risk Assessment Programme ([5]). IUCLID includes all data sets submitted by industry following Council Regulation (EEC) 793/93 on the 'Evaluation and Control of the Risks of Existing Substances'. The Regulation obliges industry to submit all readily available data on 'High Production Volume Chemicals' (HPVCs). At present, IUCLID contains 30,000 dossiers for approximately 10,500 different chemicals, and comprises the largest set of uniformly reported data for Volatile Organic Compounds (VOCs) and semi-VOCs that are directly applicable for the EU.

In accordance with Directive 67/548/EEC and Regulation (EEC) 93/793, exposure related information was to be provided for notified new chemicals and for prioritary existing chemicals, and particularly information on proposed use. When neither measured nor estimated exposure data are provided by the responsible industry (i.e. the notifier of a new chemical, which can be the manufacturer or importer of a priority existing chemical, respectively), the information on proposed use will be useful to competent authorities for developing emission scenarios. They are in most cases based on more in-depth studies of the environmental emission of chemicals used in the different industrial categories, as defined in the European Commission Technical Guidance Document on risk assessment (TGD) [8]. The emission of a chemical at different stages of its life cycle should thus be estimated by order of preference from:

1) Specific information for the given chemical (e.g. from producers, product registers or open literature)
2) Specific information from the emission scenarios document, available for several industrial categories
3) Emission factors as included in emission tables in the TGD where the emission is given as a fraction of the produced or used amount

Although information from industry is theoretically preferable, it is often the case that no such information is available. For this reason, the emission estimates in the existing RARs are obtained as a result of various possible combinations of expert judgement and empirical assumptions, based on generic scenarios defined in the TGD.

For each chemical, environmental concentrations are then calculated from emissions using mathematical models such as EUSES [9], and compared with monitored values. Emission estimates are thus a key parameter for modeling, and at the same time their estimate is complex and involves a wide range of fundamentally different parameters with varying uncertainties.

On the basis of the information submitted by manufacturers and importers, the European Commission, in consultation with Member Sates, in the past has drawn lists of priority chemicals or groups of chemicals requiring immediate attention because of their potential effects on man or the environment [e.g.10]. A risk assessment requires an extensive effort in data collection and evaluation.

As data on emissions of chemicals to specific compartments are not estimated routinely out of the risk assessment reports, a challenge can be faced when predicting emissions of other chemicals using actually available information, in a cost-effective and parsimonious way. For each chemical there are a number of parameters, such as produced amount, use and physico/chemical properties, which determine the emissions to the environment. In general, there is no obvious functional relationship between these parameters and emissions that can be derived from the methodology described in the TGD.

Data mining techniques ([11], [12]) have proven to be capable of unveiling data relationships where traditional methods, such as standard linear or nonlinear regression, perform poorly. For this reason, we attempted to estimate chemical emission volumes to environmental compartments using such techniques.

In this work focus is on (semi)Volatile Organic Compounds (VOCs), which cover a large group of chemicals predominantly found in industrial processes and in many consumer products. Some emissions are from point sources and can give rise to high local concentrations (hot-spots) in adjacent surface water, soil and the surrounding air. Other emissions are from diffuse sources.

We discuss how the extensive work that has already been done, and the data presented in the RARs, can be used to model spatial emission patterns of chemicals in a defendable way for screening level applications, by training an appropriate decision tree. Firstly, we discuss the information available in the RARs for existing chemicals and the data mining methodology used to estimate the emission of chemicals to a specific compartment from widely available information on chemical production quantities and physico-chemical properties. In the last part of the report we briefly illustrate how the method can be used in conjunction with geographic information system (GIS) processing of spatial data to build maps of emissions of chemicals at the continental scale.

2. Materials and Methods

The method aimed that we use for producing emission maps for (semi)Volatile Organic Chemicals (VOCs) comprises two types of tools, namely the data mining technique of decision trees, and geographic information systems (GIS).

Decision tree analysis ([13] , [11], [12]) is a method that analyzes (mines) a set of data and generates a decision tree that can be used to predict the value of a target variable based on the values of a set of predictor variables. Like a real tree, a decision tree has a root, branches and leaves. A prediction is made by entering the tree at the root and following the branches left or right based on values of the predictor variables until a leaf is reached. Each leaf shows the most likely value for the target variable given the set of predictor values that led to the leaf. Available software such as DTREG (www.dtreg.com) allows building decision trees automatically once having specified the predictor and target variables. This is accomplished by dichotomic search [e.g. 14] across the data set, and is a computationally intensive process. However, its simplicity from the user's perspective and the lack of need for any model assumption nor calibration make this type of methods very attractive.

V-fold cross validation [13] is performed, which is a technique for performing independent tree size tests without requiring separate test datasets and without reducing the data used to build the tree. The optimal size determined by cross validation is the best tree to use for scoring future datasets.

The basic information that is present in finalized and draft RARs includes:

1) Produced and used amount of specific chemical in the EU
2) Names (locations) of producers and importers. No link is available between company name and produced/processed amount of chemicals
3) Products and uses that the chemicals are associated with. The use is categorized as "use in closed systems", "use resulting in inclusion into or onto matrix", "non-dispersive use" and "dispersive use"
4) Emissions to wastewater, air, soil, surface water, sea/estuaries and landfills on a local scale (non-dispersive and wide-dispersive), regional and continental scale
5) Predicted Environmental Concentrations (PECs) for the same compartments and spatial scales as above
6) Physico/chemical properties
7) Ecotoxicological and human toxicological parameters

Points 1) to 4) and 6) are relevant for predicting and mapping emissions.

Data for 35 risk assessed chemicals are used as training set to build decision trees. Predictor parameters are production and use data, compiled in Table 1, and three physico/chemicals parameters in Table 3. The target parameters are emissions, compiled in Table 2. Only one target parameter can be chosen for each analysis (tree). As an example, a decision tree for the local emission to wastewater from non-dispersive emission (column 3 in Table 2) is shown in Figure 1.

Table 1 Used amounts, use in industry and consumer products of 35 selected training chemicals, ECB reports. The last row, formaldehyde, is a not-assessed test chemical where data is from EUROSTAT and SPIN. Units are in tonnes per year.

A Chemical	B CAS no	D Total annual production	E Use in closed industrial processes[1)]	F Use as intermediate[2)]	G solvent[3)]	H solid products (fixed or dissolved in matrix)	I detergents[3)]	J clothing, textiles, leather	K cosmetics	L other	M formed by natural/ industrial processes	N Wide dispersive use (sum column G to M)	O fuel (traffic)
1,4-dichlorobenzene	106-46-7	25500	100	7118	35,8	7240		0				7276	
2-(2-butoxyethoxy)ethanol	112-34-5	46600	0	2330	16800		27500					44300	
2-(2-methoxyethoxy)ethanol	111-77-3	20000	0	900	8100							8100	
2-ethylhexyl acrylate	103-11-7	70000	0	89982	16,2	1,80						18	
3,4-dichloroaniline	95-76-1	12000	0	10746								0	
4,4'-methylenedianiline	101-77-9	432000	0	432000	4000	0						4000	
4-chloro-o-cresol	1570-64-5	15000	0	14925						75		75	
acetonitrile	75-05-8	10000	9300		700						119392	120092	
acrylaldehyde	107-02-8	100000	0	100000				0		0	0	0	
acrylamide	79-06-1	100000	0			100						100	
acrylic acid	79-10-7	810000	41500	415000	1527	540		166				2233	
acrylonitrile	107-13-1	1250000	0	1249375	56,3	406		163				625	
aniline	62-53-3	530000	0	547739	760	1940				0		2811	
benzene	71-43-2	7247000	0	7247000								1410000	1410000[4)]
benzene, c10-13 alkyl derivs	67774-74-7	450000	0	278600			1400					1400	
bis(pentabromophenyl)ether	1163-19-5	0	0			6710		1500				8210	
but-2-yne-1,4-diol	110-65-6	185000	3330	183995	370							370	
buta-1,3-diene	106-99-0	1892000	0	1816308		12,0					3784	75692	71896
chloro alkanes, c10-13	85535-84-8	15000	9430		1845	1310		573		50		3778	
cumene	98-82-8	4100000	0	3310750								205000	205000
cyclohexane	110-82-7	880000	0	864000	36000	0						36000	
dibutyl phthalate	84-74-2	26000	0		3750	13500						17250	
dimethyldioctadecylammonium chloride	107-64-2	5651	468		4506			667				5173	
diphenyl ether, pentabromo deriv.	32534-81-9	0	0			1100						1100	
edetic acid (EDTA)	60-00-4	53900	6989			0	13304	3,20	756			14247	
ethyl acetoacetate	141-97-9	10000	0	9460	660							660	
methacrylic acid	79-41-4	40000	0	24276	121	30						151	
methyl acetate	79-20-9	30000	0	4800	11189		5594					16783	
naphthalene	91-20-3	200000	0	137000	2000	3352						25352	20000
pentane	109-66-0	55000	0		7825	5848						13673	
phenol	108-95-2	1829100	0	1642500								0	
propan-1-ol	71-23-8	5000	0	13550	5793		828		3310	828		16550	
styrene	100-42-5	3743000	0	3740006		1994						1994	
tetrachloroethylene	127-18-4	164000	14000	30000	1600		62400					64000	
trichloroethylene	79-01-6	138000	63140	45000	29860							29860	
formaldehyde	50-00-0	4118000	0	1400120	1812253		905960					2718213	

[1)] Non-dispersive use

[2)] Use in closed systems, or inclusion into/onto matrix

[3)] Households, professional trade etc.

[4)] Amount of benzene in petrol in Western Europe (2000)

Table 2 Measured or estimated emissions for 35 selected training chemicals, ECB reports. The last row, formaldehyde, is a not-assessed test chemical where emissions are found from decision trees. Surface water, sea/estuaries and landfills are omitted for clarity reasons.

Chemical	CAS no.	wastewater				air				soil			
		local non-disp emission[1] (kg/d)	local wide-disp emission[2] (kg/d)	regional[3] (kg/y)	EU (kg/y)	local non-disp emission[1] (kg/d)	local wide-disp emission[2] (kg/d)	regional[3] (kg/y)	EU (kg/y)	local non-disp emission[1] (kg/d)	local wide-disp emission[2] (kg/d)	regional[3] (kg/y)	EU (kg/y)
1,4-dichlorobenzene	106-46-7	5,94		45450	423450	255		782500	7258200	0,05			
2-(2-butoxyethoxy)ethanol	112-34-5	134	32,7	1941800	17483500	5,63	166	711750	6424000				
2-(2-methoxyethoxy)ethanol	111-77-3	3682		690000	25800	220		88200	12300				
2-ethylhexyl acrylate	103-11-7	701		38170	13710	33,3		10600	51780				
3,4-dichloroaniline	95-76-1	2,83		22,7	204	0,0617		0,0123	0,111	0		0	0
4,4'-methylenedianiline	101-77-9	0,78		283	2550								
4-chloro-o-cresol	1570-64-5	7,97		39750		0,658		4650		0,006		1275	
acetonitrile	75-05-8	2777	0	31201	4496800	1528		10060400	99954290				
acrylaldehyde	107-02-8	20,4		6205	62050	0,287		657000	15768000				
acrylamide	79-06-1	0,502		9120	84000	0,200		66	103				
acrylic acid	79-10-7	323	0,335	218000	973000	21,9	0,205	54000	277000				
acrylonitrile	107-13-1	22,2				330		330000	3310000				
aniline	62-53-3	6,19		260	2300	10,3		16000	146000				
benzene	71-43-2	903		2585000	23262000	1413		18291000	165000000	362		65500	590000
benzene, c10-13 alkyl derivs	67774-74-7		0,15	108040	2000200								
bis(pentabromophenyl)ether	1163-19-5	4,84	0,2	121740	334860	0,17		2909	26150			14800	106200
but-2-yne-1,4-diol	110-65-6	20,7	11	58040	580400	0,167	0,29	8,7	87	0		0	0
buta-1,3-diene	106-99-0	476		120240	1074490	3750		1435300	12496400				
chloro alkanes, c10-13	85535-84-8	52,3		174102	1738799	0,512		39,4	394				
cumene	98-82-8	2500		615000	6150000	2183		1242300	12423600	33,3		8190	81900
cyclohexane	110-82-7	333	12	736400	6625600	6625600	6625600	6895000	56893000	0			
dibutyl phthalate	84-74-2	30,4		94590	610860	18,1		110865	403599				
dimethyldioctadecylammonium chloride	107-64-2	26,2	0,507	9600	86000	0		0	0			11000	99000
diphenyl ether, pentabromo deriv.	32534-81-9	0,15		44,6	136	0,124		4339	38831			1590	14270
edetic acid (EDTA)	60-00-4	561	2	2895000	26059000	0,0242						582000	5239000
ethyl acetoacetate	141-97-9	71,1	25,6	1000	10200	2,33	161	5000	49400				
methacrylic acid	79-41-4	1086		80000	325000	3333		4000	37000			0	0
methyl acetate	79-20-9	313	3,55	45000	402000	14618	17876	1328000	11958000				
naphthalene	91-20-3	21,4		12077	46008	16,1		3608906	32119658	0		9138	40368
pentane	109-66-0	1	0,0371	6810		9101	14,2	3744977					
phenol	108-95-2	82,5	0,96	513000	4618000	48,5		9683000	87146500	0		600	5100
propan-1-ol	71-23-8	105		1516858	3008142	2050		2104827	4176873				
styrene	100-42-5	839		252000	1038030	1836		2010000	16944000	0		0	0
tetrachloroethylene	127-18-4	1,29		13281	119298	791		9936000	44834400				
trichloroethylene	79-01-6	13,0	0,096	522467	4660587	3451	19,9	6373080	48196910	2,44		7300	60400
formaldehyde	50-00-0	2500	32,7	284022	2574994	4142	148	877832	8215811	3,58		9184	83103

[1] For local emissions all the non-dispersive emission amounts are added. This corresponds with the approach of applying local emission amounts to arbitrary sites, in order to find potential risk sites/scenarios. When adding all emissions the worst-case scenario is defined, i.e. a site where all local emissions are arbitrarily situated at the same site, and thus releasing to the same recipient (wasterwaterplant, river, air, soil etc.)

[2] Often the local emissions stated in the RAs only comprise non-dispersive sources, i.e. industries etc. The wide-dispersive household uses are often not included in the local emission, but only in the regional and continental emissions

[3] All emissions from both non-dispersive and wide-dispersive sources are considered in the determination of a regional background emissions

Table 3 Physico/chemical parameters for 35 training chemicals. The last row, formaldehyde, is a not-assessed test chemical. Vapour pressure (vp), Octanol-water partitioning coefficient (logKow), Water solubility (watsol) are from IUCLID.

Chemical	CAS no.	vp (kPa)	logKow	watsol (mg/l)
1,4-dichlorobenzene	106-46-7	160	3,38	65
2-(2-butoxyethoxy)ethanol	112-34-5	0,0027	0,56	1000000
2-(2-methoxyethoxy)ethanol	111-77-3	0,03	-0,682	1000000
2-ethylhexyl acrylate	103-11-7	0,0155	3,89	9,6
3,4-dichloroaniline	95-76-1	0,000184	2,7	580
4,4'-methylenedianiline	101-77-9	2,87E-09	1,59	1,25
4-chloro-o-cresol	1570-64-5	0,02666	3,09	2300
acetonitrile	75-05-8	9,864	-0,34	139000
acrylaldehyde	107-02-8	29,3	-0,89	240000
acrylamide	79-06-1	0,0009	-1	2155
acrylic acid	79-10-7	0,38	0,46	1000000
acrylonitrile	107-13-1	11,5	0,25	0,735
aniline	62-53-3	0,04	0,9	35000
benzene	71-43-2	9,97	2,13	1800
benzene, c10-13 alkyl derivs	67774-74-7	0,0013	8,31	0,041
bis(pentabromophenyl)ether	1163-19-5	4,63E-09	6,27	0,0001
but-2-yne-1,4-diol	110-65-6	0,00017	0,73	0,75
buta-1,3-diene	106-99-0	240	1,99	735
chloro alkanes, c10-13	85535-84-8	2,13E-05	6	0,47
cumene	98-82-8	0,496	3,55	50
cyclohexane	110-82-7	10,3	3,44	58
dibutyl phthalate	84-74-2	0,97	4,57	10
dimethyldioctadecylammonium chloride	107-64-2	0	3,8	2,7
diphenyl ether, pentabromo deriv.	32534-81-9	4,69E-08	6,57	0,0024
edetic acid (EDTA)	60-00-4		-5,01	400
ethyl acetoacetate	141-97-9	0,1	0,25	125000
methacrylic acid	79-41-4	0,09	0,93	89000
methyl acetate	79-20-9	21,7	0,18	272000
naphthalene	91-20-3	0,0072	3,55	30
pentane	109-66-0	56,58	3,45	38,5
phenol	108-95-2	0,02	1,47	84000
propan-1-ol	71-23-8	1,94	0,34	1000000
styrene	100-42-5	0,667	3,02	300
tetrachloroethylene	127-18-4	1,9	2,53	149
trichloroethylene	79-01-6	8,6	2,29	1100
formaldehyde	50-00-0	0,75	-0,78	550000

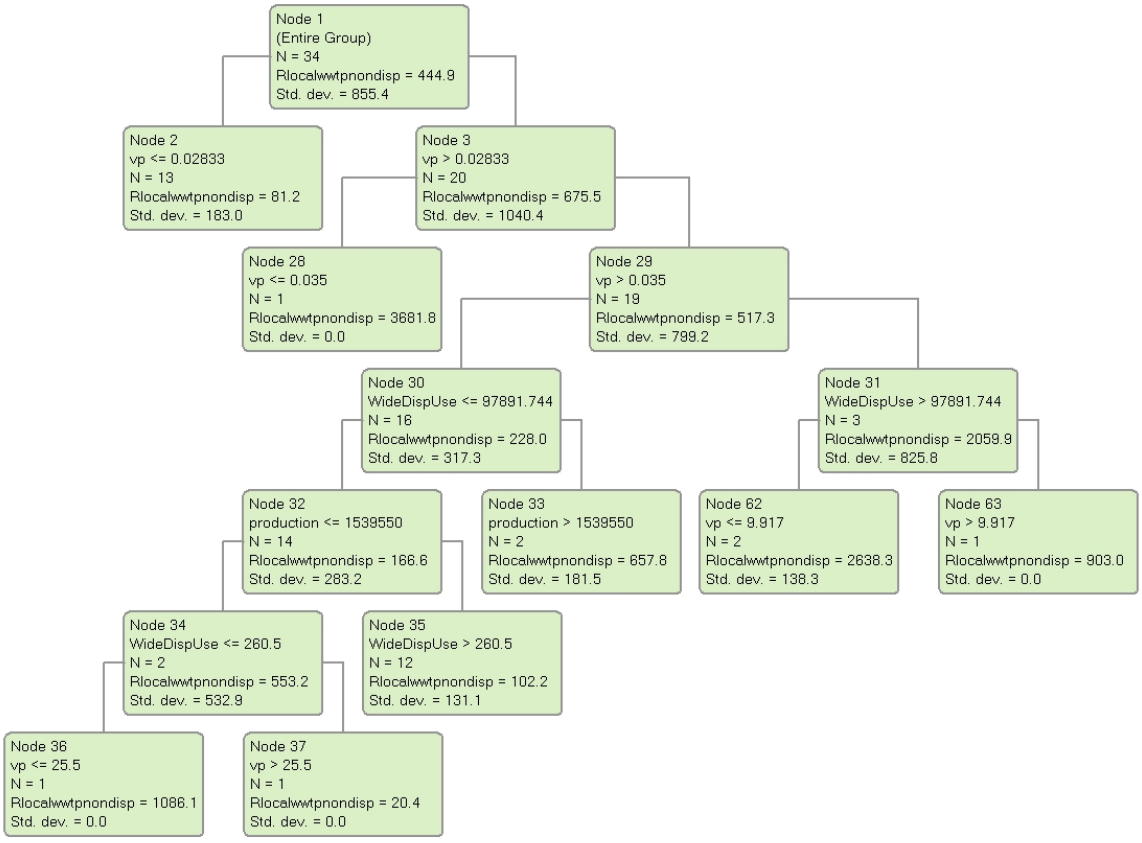
Decision trees can be built for all forms of emissions to environmental compartments. The training set is used for cross-validating the emissions for each chemical by removing it from the training set and building a decision tree with the remaining chemicals. Figure 2 shows the cross-validation in a scatter diagram between reported emissions and predicted emissions. As it is shown, the decision tree allows conservative estimates where 69% of the predictions are overestimating the emissions relative to reported values and that 51% of the predictions are within one order of magnitude (±) of the reported values. Conservative estimates are typically occurring for low reported emissions.

The method we propose for predicting emissions for not assessed chemicals consists of the following steps:

1) Retrieve gross production and consumption data and physico-chemical properties for a VOC of interest
2) Apportion the total consumption volume of the chemical to different use modes
3) Identify a mode of emission of the chemical to the environment, e.g. in treated wastewater, direct to air

4) Using the data on consumption volume, use modes and physico-chemical properties of the substance as predictor parameters, enter a decision tree built from the training set to estimate the selected mode of emission (target parameter); this provides an emission total with reference to the area of origin
5) Apportion the emission total to point sources (emissions from individual high production volume plants) and to diffuse emissions (dispersive use in industrial areas, and in households);
6) Construction of emission maps by summing the point emissions to a map of distributed diffuse emissions, which can be evaluated from total diffuse emission using an emission pattern such as land use or population density as explained in the following.

Figure 1 Decision-tree where production, product, use (cf. Table 1) and physico/chemical parameters (cf. Table 3) for 35 different VOCs and semiVOCs are used as predictor parameters. The target parameter is local emissions to wastewater from non-dispersive sources (column 3, Table 2). In each node the needed split parameter value is shown, number of attributed chemicals, emission to wastewater from local non-dispersive use (target parameter) in kg per day, and the emission standard deviation.



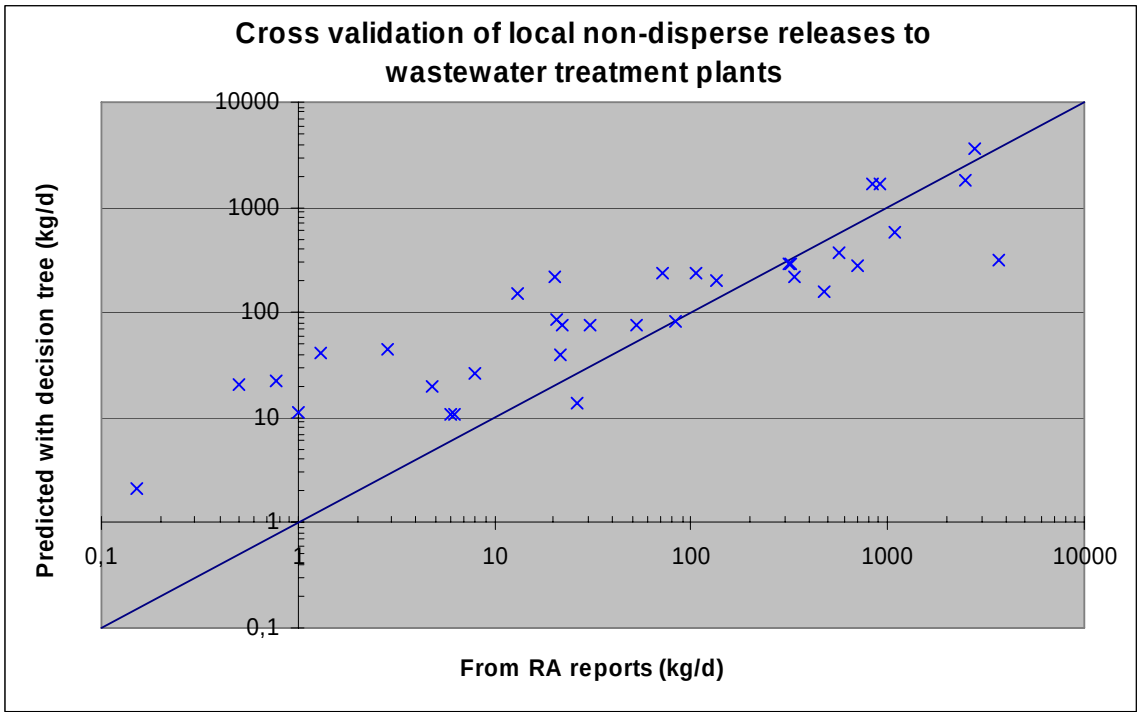
In order to accomplish the above stated steps, one should use the best data available in the specific situation to which the estimate is referred. Here we focus on data available for estimates at the level of continental Europe.

The total annual use can be computed from the mass balance:

Quantity used = quantity produced + quantity imported – quantity exported

Production, import and export data are found in Eurostat, either on a national basis or on EU scale.

Figure 2 Cross-validation of predicted local non-dispersive emission to wastewater (Rlocalwwtpnondisp) for 35 training chemicals (no local emission is estimated for benzene, c10-13, alkyl derivatives). A decision tree is built for each chemical by omitting the chemical in the building of the tree.



Use mode information can be found from the database of Substances in Preparations in Nordic Countries (SPIN) [15], which provides data on the use of chemical substances in Norway, Sweden, Denmark and Finland. Conditions in the Nordic countries are thus extrapolated to the entire EU, although some differences in chemicals composition of products, use patterns of products and emission factors of chemical consuming industrial processes may prevail. The SPIN database has a degree of detail that can not be found in other databases in the EU, so this assumption can be considered to be realistic and appropriate as a first approach.

Finally, physico-chemical parameters can be found from a variety of databases. In this analysis, the IUCLID database is used. In this way, a compilation of approximate information can be retrieved for a large number of existing VOCs.

3. Results and Discussion

3.1 Estimate emissions to environmental compartments for non-assessed chemicals

One way to use decision tree analysis, as shown above, is to predict emissions of chemicals for which a risk assessment report, or generally speaking comprehensive information on emissions,

is not available. As a test chemical that is used in large amounts and in a variety of activities and products formaldehyde, CAS no. 50-00-0, is chosen.
From Eurostat the formaldehyde production, import and export figures for EU25 for 2004 are extracted, yielding a used quantity of 4118 ktonnes per year
From the Nordic SPIN database information on categories “Industrial use” and “Use category” are extracted. This is used to assign the 4118 ktonnes per year relatively in categories defined in the Common Reporting Format (CRF) categories for “solvent and other product use”, as defined in the IPCC Guidelines for National Greenhouse Gas Inventories [17]; “paint application”: 8.08e-05, “degreasing and dry cleaning”: 0.22, “chemical products manufacturing and processing”: 0.34, “other”: 0.44. “Solvent and other product use” does not cover every aspect of a (semi)VOC emission since, for instance, energy and transport may be significant sectors for some chemicals such as benzene. However, for most (semi)VOCs the main uses are within this category.
For formaldehyde predictor parameters analogous to training chemicals are found. Last row in Table 1 is found by distributing the total annual use in the EU in the following way: “Paint application” is assigned to column G (solvent), “degreasing and dry cleaning” to column I (detergents), “chemical products” to column F (use as intermediate) and “other” to column G (solvent) since the main part of “other” is as conserving agent. Last row in Table 3 is made from IUCLID data. Emissions (target parameters) are estimated from decision trees built from the training set, and shown in last row in Table 2.

3.2 Emission mapping with GIS

In the previous section it is shown how local and wide-dispersive emissions of chemicals, from use in industries and consumer products, are estimated. The local emissions are generated at industrial sites where chemicals are produced or products are processed, and wide dispersive (regional and EU) emissions occur in industrial, urban or agricultural areas where the chemicals are used.

Confidentiality in RARs prevents direct allocation of industrial emissions. These can be assigned, as a proxy, to locations representing large chemical industrial plants, which are covered by the European Pollutant Emission Register (EPER) [16]. EPER covers 21 very general classes of industrial activities. Emissions to air, direct and indirect to water are reported from approximately 10,000 large and medium-sized industrial facilities in the 15 EU Member States, Hungary and Norway. EPER has information on 50 single chemicals or groups of which six single chemicals are included in the ECB priority chemicals. Some ECB priority chemicals are included in EPER as chemical groups, e.g. phenols, cyanides and non-methane VOCs.

The share of emissions from all sources covered by EPER inevitably varies for each Member State, industrial activity and pollutant. For some air pollutants the EPER share can be assessed, whereas for direct and indirect emissions to water this is more difficult due to a lack of pan-European data sets. As an example, a comparison with the EU15 total emissions of some important greenhouse gases and air pollutants (as reported under the UN Framework Convention on Climate Change and the UNECE Convention on Long-Range Transboundary Air Pollution) shows that EPER covers around 6% of EU15 total Non-Methane VOC emissions. This underestimation for more than a factor of 10 can be due to the fact that the RAR emissions also cover households and wide-spread product use, whereas EPER only covers industrial activities. It excludes for example emissions from the transport sector and from most agricultural sources,

whereas the underlying totals include these emissions. For these reasons, it seems advisable to take from the EPER inventory only relevant information on the potential location of industrial plants.

The main industrial activities contributing to chemical emissions, according to the RARs, are production, processing and formulation. These activities usually can be attributed to the EPER activity “Basic organic chemicals”. The locations of the EPER-reported “Basic Organic Chemicals” facilities are displayed in Figure 3.

The names of the facilities associated with this activity in EPER for a given chemical do not completely correspond to the facilities stated in the RARs. This can be due the fact that the RARs are typically elaborated in the mid 1990s, whereas EPER is up to date. However, the *number* of production and importing facilities in RARs are in reasonable agreement with the number of “basic organic chemicals” facilities in EPER. This applies both to the number of facilities per country and the total number of facilities in the EU. The reason for a slightly higher number in RARs could be the inclusion of smaller downstream processing plants.

In order to use the updated facility information in EPER together with the in-depth RAR emission estimates, the following assumptions are made with respect to mapping non-dispersive emissions from industrial point sources:

a. EPER category “Basic organic chemicals” facility sites are used for EPER as well as non-EPER chemicals (EPER comprises only a sub-set of the chemicals forming a subject of RARs)

b. RAR categories “Use in closed systems”, “Use resulting in inclusion into or onto matrix” and “Non-dispersive use” are related to EPER “Basic organic chemicals” facilities

c. RAR emissions are attributed to randomly selected EPER “Basic organic chemistry” facilities. This is done for the number of sites stated in the RARs for each chemical

d. Non-dispersive emissions associated with other industries than “Basic organic chemicals” should be located as such, if information is available from EPER.

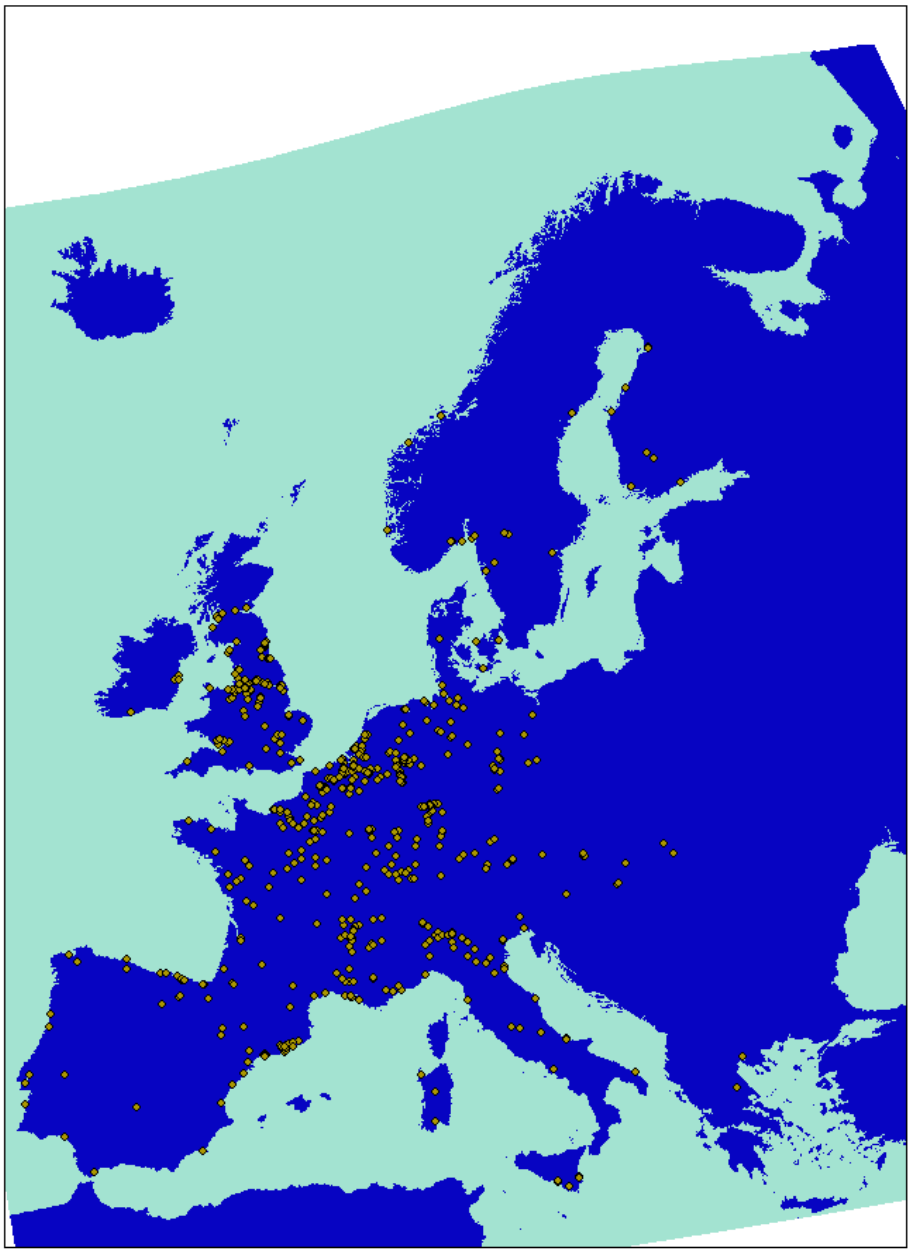
Wide-dispersive uses are related to public/household use and to other use delivering uncontrolled exposure, such as disperse industrial activities and traffic. Public use can be either outdoor or indoor, and for some chemicals this distinction can be made based on use patterns of specific products groups, such as detergents, cosmetics, disinfectants, household paints. Emissions attribution to maps is summarized in Table 4.

For urban and industrial areas, the following algorithm has been adopted in order to represent the emissions E at location (x,y) (in units of [M] [L]-2[T]-1) from knowledge of the fraction of area occupied by urban or industrial land uses, LU, at the same location:

$$E(x,y) = \frac{E^* LU(x,y)}{\int_A LU(x,y)dxdy}$$

E* being the emission over area A (in units of [M][T]-1). Area A can be the entire continent, or a sub area depending on the resolution on aggregated emissions available (for instance, at the national or province level). This is equivalent to assign emissions to a grid cell in a map as proportional to the share of the land use considered as a source of emission.

Figure 3 – “Basic organic chemicals” facilities from EPER



Population density is mapped by many organizations across the world; recently, a particularly upgraded product seems to be the Gridded Population of the World (GPW: http://www.ciesin.org/datasets/gpw/globldem.doc.html). Many algorithms can be used in order to obtain an emission map. In our case, we propose the above where LU(x,y) is replaced by the population density map D(x,y). From the above equation, it is clear that emissions related to population density are all similar to each other, apart from a scaling constant given by the total emission E*.

Table 4 Summary map-key to where the emissions are attributed

	Local non-dispersive emission (kg/d)	Local wide-dispersive emission (kg/d)	Regional emission[1)] (kg/y)
Chemical industries	EPER "Basic organic chemicals"		Population density or CORINE land cover (urban categories)
Other industrial activities	EPER site or CORINE land cover (industrial/commercial land uses)	included in regional emission	Population density or CORINE land cover (industrial/commercial land uses)
Households		included in regional emission	Population density
Urban areas and professional workers		included in regional emission	CORINE land cover (urban categories)
Transportation		included in regional emission	Traffic density

Blank spaces: not relevant

[1)] The regional emissions calculated in the RARs are thus defined for standard densely populated and highly industrialised areas

To illustrate the differences in use amounts, product types, use patterns, spatial scales and emission patterns for industrial chemicals, three risk assessed chemicals and formaldehyde are selected for mapping the releases to wastewater. Non-dispersive emissions from industrial sites are attributed with randomly selected EPER sites in consistence with the national total number of sites from the RARs.

(1) Benzene is used in much larger amounts and is predominantly used as intermediate in chemical processing of a variety of other chemicals. The emission is related to the chemical processing and is thus non-dispersive primarily to air and wastewater, but significant emissions to soil also occur at a local scale in the vicinity to the industrial sites. The most important emission source for benzene is, however, through the use of fossil fuel where benzene is a natural component of crude oil, and therefore an intrinsic constituent of certain refinery fractions. The most significant amount of benzene is found in motor fuel, with concentrations of 1-5%. This emission is wide-dispersive and can be correlated with traffic intensity but since traffic density maps are not available urban land use pattern is used.

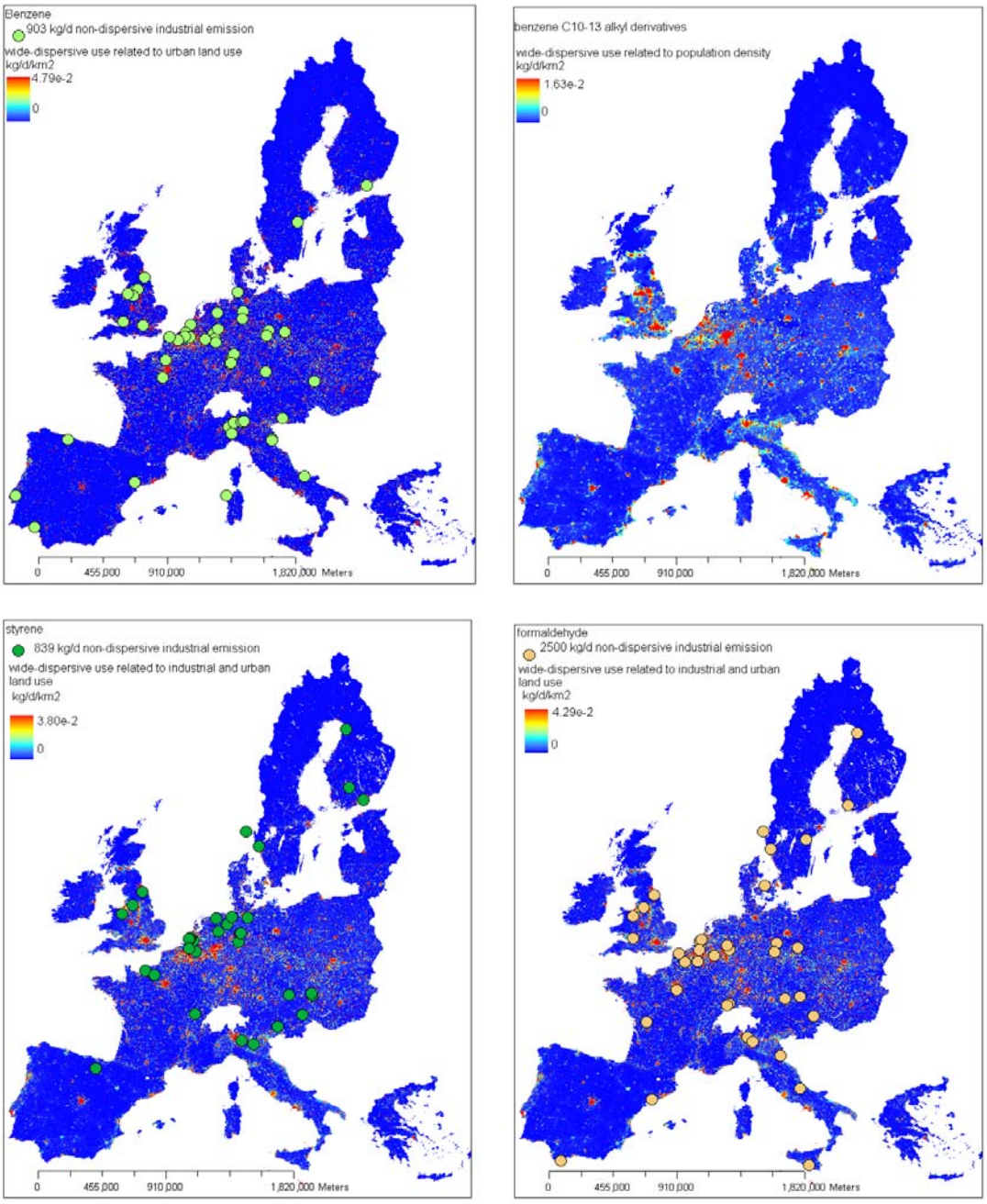
(2) C10-13 alkyl derivatives of benzene are only used as intermediates in the production of linear alkylbenzene sulphonate (LAS), which is used as a household detergent, and is emitted to wastewater. The use is wide-dispersive only and the emission pattern is given by the population density pattern.

(3) Styrene is a monomer exclusively used in the production of various polymer products, such as polystyrene. Contrary to the previous two chemicals styrene is incorporated in a solid matrix in products used for packaging, building, transport and even clothing. The content of styrene residuals in polymers and copolymers is approximately 0,04% of the used monomer amount. Emissions are predominantly associated with the local scale in the vicinity of the processing industries, where high emissions to air and wastewater can be found. The wide-dispersive emission pattern for styrene is given by the superposition of industrial land use, with urban land use pattern. The two patterns have the same weight, as we assumed the regional emissions were half to each category.

(4) Formaldehyde is considered as an example of non-risk assessed chemical. The wide-dispersive emission pattern for formaldehyde is given by the superposition of industrial land use, with urban land use pattern. The two patterns have the same weight, as we assumed the regional emissions were half to each category.

Non-dispersive point emissions are superimposed with wide-dispersive background emissions. Local wide-dispersive emissions are included in the regional and EU emissions. Figure 4 shows the maps of emissions.

Figure 4 Emission pattern to wastewater for a) benzene, E*= 7082 kg/d; E_max= 4.79E-02 kg/d/km^2, b) benzene derivatives, E*= 296 kg/d; E_max= 1.63E-02 kg/d/km^2, c) styrene, E*=690 kg/d; E_max= 3.80E-02 kg/d/km^2, d) formaldehyde, E*=778 kg/d; E_max= 4.29E-02 kg/d/km^2.



In summary, we used emission data available for 35 risk-assessed chemicals as a training set for predicting and mapping emissions of chemicals that have not been assessed, but may pose a risk to humans and the environment. Emission predictions have been done to air, wastewater, soil, and can be extended to inland surface water and the sea. The method of decision trees, a data mining technique for predicting a target value based on a set of predictor variables, generates

emission estimates which have been cross validated and shown to bear acceptable error. In particular, emissions modeled according to this approach are within one order of magnitude with respect to RAR data.
The emission model that has been developed requires data on use amounts in industry and in downstream products (that can be retrieved e.g. from Eurostat and the Nordic SPIN products database) and vapor pressure, logKow and water solubility (that can be retireved e.g. from the IUCLID database), of the test chemicals. This information is readily available for many chemicals in a transparent and uniformly comparable form. For this reason, the method can be used at a screening level to map emissions for chemicals not subjected to Risk Assessment Reports, for the goal of chemical fate and transport modeling and spatially distributed evaluation. The procedure can be also seen as supportive to the development of a RAR, as it allows simplification and acceleration of the process of emission estimates, which can be a very time consuming task.
As a screening level procedure, it provides only a first approximation estimate, although it is observed that often emission inventories themselves have intrinsically high uncertainties ([18]); therefore, the proposed procedure appears promising when only limited data are available and a quick response is required. Also, it should be noticed that when pursuing assessments at local scale, the spatial distribution of point emissions, which are selected at random from the EPER inventory, may have a very strong impact on predictions. In general, however, the procedure is expected to provide more and more reasonable estimates as one moves to scales such as a region or the continent.

4. Acknowledgements

This research was financially partly supported by the European Union under European Commission FP6 Contract No. 003956 (NoMiracle Project). We wish to thank all colleagues from NoMiracle RP 1 that helped with discussion the development of the work, and we dedicate it to the memory of Kirsten Voormann, who passed away prematurely during the development of the project.

5. References

1. Pärt, P. (main author), Environment and health, EEA Report No 10/2005, Copenhagen, 2005
2. Karjalainen,T., Commission research in Action : tackling the hormone disrupting chemicals issue, EUR report 21941, 2005
3. UNECE/EMEP Task Force on Emissions Inventories and Projections, EMEP/CORINAIR Emission Inventory Guidebook – 3rd edition, EEA Technical report No 30, Copenhagen, 2005
4. UNECE Convention on Long-range Transboundary Air Pollution, http://www.unece.org/env/lrtap/full%20text/1979.CLRTAP.e.pdf
5. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), Risk assessment Reports, http://ecb.jrc.it/ASSESSMENT_OF_CHEMICALS/ , various years
6. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), European INventory of Existing Commercial chemical Substances (EINECS) http://ecb.jrc.it/esis/esis.php?PGM=ein

7. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), The International Uniform ChemicaL Information Database (IUCLID) http://ecb.jrc.it/iuclid4/
8. EC, 2003, Technical Guidance Document in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. (http://ecb.jrc.it/tgd/)
9. EC (2004) European Union System for the Evaluation of Substances 2.0 (EUSES 2.0). Prepared for the European Chemicals Bureau by the National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands (RIVM Report no. 01900005) (http://ecb.jrc.it)
10. European Commission, DG JRC, European Chemicals Bureau (ECB) (coordination), Priority Lists, http://ecb.jrc.it/priority-setting/
11. Breiman, L. , J. Friedman, R. A. Olshen and C. J. Stone, "Classification and regression trees". Pacific Grove, Wadsworth, 1984.
12. Berikov, V., A.Litvinenko, "Methods for statistical data analysis with decision trees". Novosibirsk, Sobolev Institute of Mathematics, 2003 (http://www.math.nsc.ru/AP/datamine/eng/decisiontree.htm).
13. Sherrod, P.H., DTREG - Classification and regression trees and support vector machines for predictive modeling and forecasting, User Manual, 2003
14. Gillies, D., Artificial Intelligence and Scientific Method, Oxford: Oxford University Press, Pp. xii + 176, 1996.
15. Nordic Council of Ministers, Chemical group, Substances in Preparations in Nordic Countries (SPIN) Database, http://www.spin2000.net/spin.html
16. European Environment Agency (EEA), European Pollutant Emission Register (EPER) http://www.eper.cec.eu.int/eper/
17. Penman, J., Kruger, D., Galbally, I., Hiraishi, T., Nyenzi, B., Emmanuel, S., Buendia, L., Hoppaus, R., Martinsen, T., Meijer, J., Miwa, K. & Tanabe, K.(eds), 2001. Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories. IPCC National Greenhouse Gas Inventories Programme.
18. Breivik, K., Vestreng, V., Rozovskaya, O., Pacyna, J.M., Atmospheric emissions of some POPs iN Europe: a discussion of existing inventories and data needs (2006), Environmental Science and Policy, 9: 663-674.

Annex – regression analysis

A decision-tree based method for estimating emissions and producing emission maps for VOCs and semi-VOCs has been discussed in this report.
This method is alternative to other techniques, such as correlation or pattern recognition followed by linear regression models, including Principal Component Regression (PCR), Partial Least Square Regression (PLS-R) and Multiple Linear Regression (MLR).
RAR data for 35 risk-assessed chemicals are used here as training set for calibration and cross-validation of regression models based on PCR, PLS-R and MLR. Predictor parameters (PP), i.e. parameters that are entered in the models, are produced amount, non-dispersive use, use as intermediate (sum of "use in closed systems" and "use resulting in inclusion into or onto matrix") and wide dispersive use, assigned by the letters D, E, F and N and compiled in Table 1 of the report, together with the physico-chemical parameters logH and logKow given in Table 3. The target emissions (TE) that are used to calibrate, the models, describe worst-case local and regional conditions, represented by local non-dispersive emissions and wide-dispersive emissions, respectively. Data on emissions to wastewater, air and soil at local and regional conditioned scenarios are compiled in Table 2.

Prior to selecting PPs for optimal modeling of TEs, the PPs and TEs were log-transformed to approximate normal distribution of data. Normal distribution can be assumed when skewness $< \pm$ 2*standard error of skewness and kurtosis < 2*standard error kurtosis [19]. The normal distribution criteria are met for all selected parameters in Table 1, all local non-dispersive emissions, regional emissions to wastewater and soil, and logKow. Parameters logH and regional emission to air showed slightly skewed distributions by having longer right tail and left tail, respectively, than those of a normal distribution [20]. The latter are nonetheless included in the analysis. Furthermore, parameters have been auto scaled, i.e. mean subtracted and divided by standard deviation, to obtain equal variances and mean zero, and approximate homoscedastic noise between variables [20, 21].

Regression models can be made for all forms of emissions to environmental compartments. In this report we find the coefficients αj to the multiple regression models, as defined in Equation 1, using a maximum of six PPs to explain the six TEs.

$$TE = \alpha 0 + \alpha 1*(\text{production, D}) + \alpha 2*(\text{non-dispersive use, E}) + \alpha 3*(\text{use as intermediate, F}) + \alpha 4*(\text{wide dispersive use, N}) + \alpha 5*(\text{logKow}) + \alpha 6*(\text{logH})$$

where TE represents the target emission, TE1 to TE6.

PCR and PLS were compared to reveal the inherent amount of correlation and co-linearity between PP and TE. Whereas PCR represents the inherent correlation, i.e. without fitting patterns in X, including all the PPs, to correlate optimal with TE data, PLS is an iterative process where the maximum amount of variation in X fitting optimal to the pattern in TE is found. PCR consists of two steps, where the first step is a PCA carried out on X after which the principal components (PCs) are used as predictors in an MLR. PLS-R is a bilinear modeling approach, where the PPs are projected onto a small number of underlying latent variables in an iterative

process. In PLS-R, the TE data are used actively in determining the latent variables ensuring highest possible relevance for prediction of TE in first PC. The number of PCs increases until no further increase in the explained TE-variance is achievable; i.e. maximum explained variance is obtained and further inclusion of PCs increases the noise in the model.

An important assumption for the MLR method is that the PPs are linearly independent. Optimal PPs with highest explanatory capacity were selected based on the results for PCR and PLS in parallel to stepwise linear regression [19, 22]. The best simple MLR results are presented together with a visualization of latent variables in loading plots from the PLS-R models.

Multiple linear regression models for the six TEs have been derived and the coefficients, α to Equation 1, are shown in Table A1. The PPs were selected by a stepwise regression procedure in SYSTAT. In parallel, PCR and PLS models were used for selection of PPs based on the weighted criteria: 1) maximum orthogonality and 2) highest explanatory capacity as discussed below in relation to Figure A1.

The usual limit used in the interpretation of a p-value is 0.05 (or 5% significance level). As observed from Table A1, the p-value is below 0.05 in all six models except for TE1, which represents local non-dispersive emissions to wastewater.

Use in closed industrial processes, E, which can be considered to be a point source, was tested as PP for local emissions, but as the number of observations was low, i.e. n = 2, 9 and 8 out of a total of n = 35, E was excluded from the regression analysis. A pair wise Pearson correlation matrix showed similar correlations coefficients of E to F and D, respectively, of 0.6. Furthermore, the pearson correlation to wastewater emissions TE1 and TE4 are highest when compared to the other TEs. Correlation coefficients for LR of E versus TE3 and TE4 were 0.01 and 0.06, respectively.

PCR and PLS were used for supporting the selection of PPs in the optimized MLR models shown in Table 1. The explanatory capacity and correlation patterns between PPs and TEs for the first two principal components of the PLS-R models is visualized in Figure A1 The loadings of PPs with respect to TE1 to TE6 (from the upper left towards the lower right plot) shows the importance of each PP in the principal components, i.e. PC1 and PC2, with respect to the X-variance in each PC used for explaining Y, i.e. the individual TEs. The used X-variance in PC1 and PC2 for explaining the TE-variance by the first two principal components are given in percent below each correlation loading plot in Figure A1. The loading plots show the associations within PP and TE. However, non-dispersive use (E) shows high correlation to target emissions for non-dispersive and wide-dispersive wastewater emissions (cf. Table XX). Therefore, E has been used.

Wastewater emissions

In general, non-dispersive and wide-dispersive wastewater emissions, TE1 and TE4 respectively, are poorly described by the models; only 19% TE-variance is described by 34% and 42% PP-variance for TE1 and TE4, respectively. Both models are one component PLS models, and TE1 and TE4 are being positioned within the inner ellipse representing less than 50% explained variance by the models. When running PCR, which is a two-step method performing a PCA on the PPs and secondary using the PCs as predictors in a MLR, the models predict 0% TE1 and TE4-variance in the first PC. The PCR reveals that the PPs included in the analysis are poor descriptors for wastewater releases. From the PLS loadings weight above, the best possible fit of TE1 and PP is suggested to be logH and N, which are close to orthogonal to each other and accordingly the best MLR model is based on these two PPs. Likewise, the PLS loading plot of TE4 and PPs shows that logKow has close to zero explanatory capacity in PC1 and N little

explanatory capacity in PC2. This means that logKow and N are close to orthogonal. Inclusion of LogKow in the MLR model for estimating TE4 results in the best MLR model. The original parameters are not successful PPs for wastewater releases as is be seen from Table A1; the wide-dispersive model diagnostics shows the best performance compared to local non-dispersive wastewater release.In spite of on the relatively high pearson correlations coefficients, simple correlations of E versus TE1 and TE4, respectively, showed poor model performance; i.e. high p-values, low F-values and lower R^2 compared to Q^2 reflecting low robustness of the model.

Table A1 MLR models of target emissions, TE, where α0 to α6 are regression coefficients, n is number of cases, R2 and Q2 are correlation coefficients based on calibration and leave-one-out cross-validation, respectively. The F-ratio is regression sum of squares divided by unexplained residual variance, p-value is the significance level for the modeled variation to be real, RMSEC is the root mean square of calibration, and RMSEP the root mean square error of predictions, expressed in the same units as TE. Bold figures indicate good combination of high n, high F-ratio and low p-value.

	Constant	Production, D	Use in closed processes, E	Use as intermediate, F	Wide dispersive use, N	logH	logKow	Model performance parameters						
Regression coefficients	a0	a1	a2	a3	a4	a5	a6	n	R2	Q2	F-ratio	p-value	RMSEP	RMSEC
Local non-disp wastewater, TE1	0.044	-	-	-	0.463	0.534	-	30	0.155	0.02	2.474	0.103	1.03	0.93
Local non-disp air, TE2	-0.01	-	-	-	-	0.696	-	29	0.396	0.341	17.72	3E-04	0.75	0.7
Local non-disp soil, TE3	-0.28	0.697	-	-	-	-	-	5	0.941	0.853	47.48	0.006	0.345	0.218
Wide-disp wastewater, TE4	3.372	-	-	-	0.506	-	-0.09	30	0.21	0.07	3.658	0.039	1.172	1.054
Wide-disp air, TE5	0.063	-	-	-	0.565	-	-0.12	28	0.533	0.398	14.27	1E-04	0.638	0.554
Wide-disp soil, TE6	-0.24	-	-	-	0.438	-	-	8	0.741	0.568	17.17	0.006	0.387	0.298

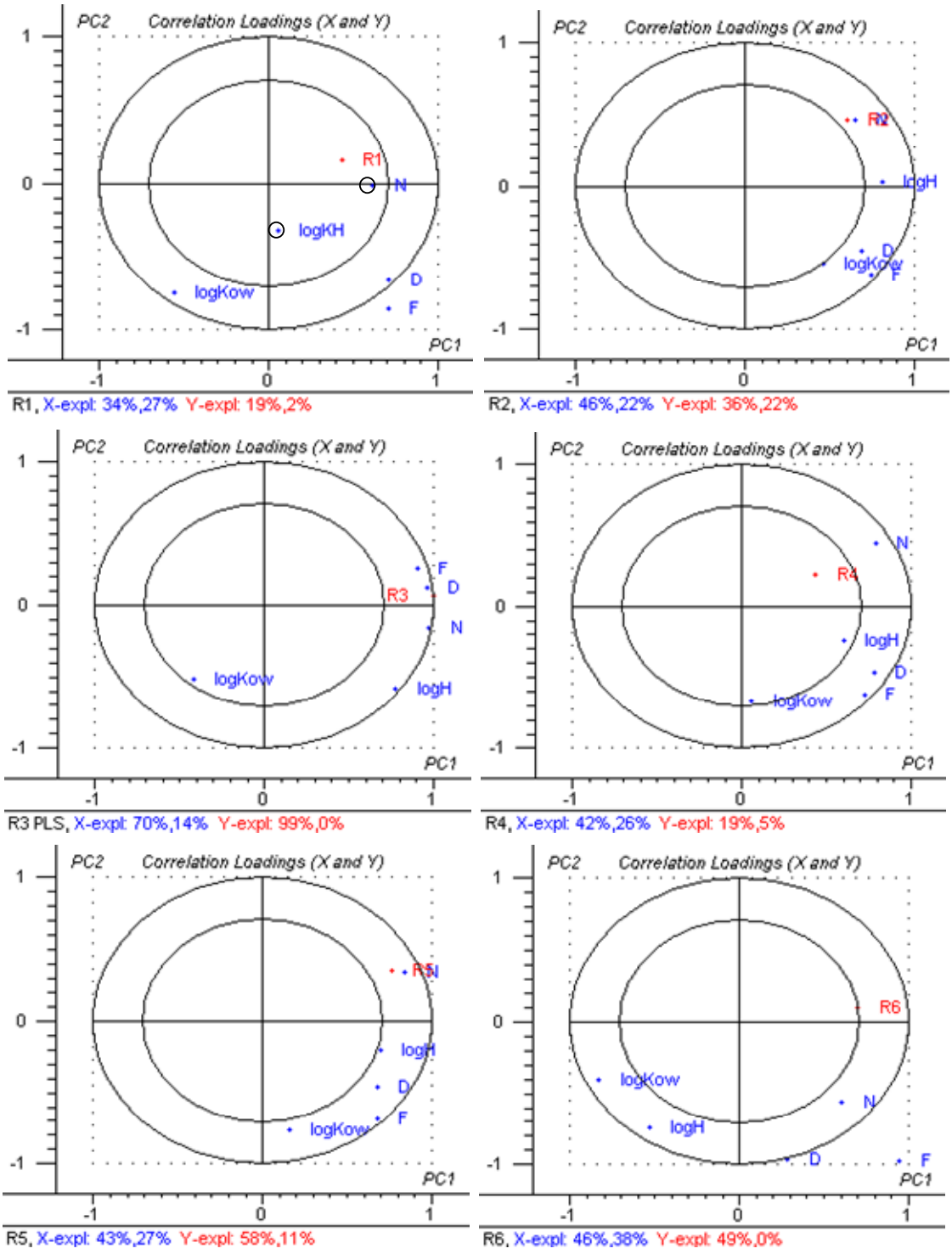


Figure A1 PP and TE loading plots for Partial Least Square (PLS) models using the original PPs; Production (D), non-dispersive use (F), wide dispersive use (N), logKow and LogH, as PPs for predicting the target emissions TE1 to TE6, as described in Table A1. TE1 to TE3 are local scale emissions to wastewater, air and soil, while TE4 to TE6 represent wide-dispersive emissions to wastewater, air and soil, respectively. The outer ellipse indicates 100% and the inner ellipse indicates 50% of explained variance.

Air emissions

The models for predicting non-dispersive (TE2) and wide-dispersive (TE5) air emissions are shown in the upper right and lower left loading plots in Figure A1. The PPs show a high degree of inter-correlation when fitted to TE2 and TE5, respectively. Furthermore, all PPs have high explanatory capacity compared to the remaining models as all PPs lie outside the 50% explained variance ellipse. When performing PCR and comparing to the PLS, the within PP and PP-TE correlation patterns do not change significantly, which reveals the PPs fit TEs well without an iterative fitting process of co-occurring variations in PP and TE; i.e. there is less noise in the air emission models compared to the wastewater release models. Due to the high degree of co-variation between PPs, the optimal LR modelling of TE2 and TE5 are univariate simple regression models. The best fit and F-statistics was obtained by using LogH as predictor for local non-dispersive emissions to air (TE2) and wide-dispersive use (N) as predictor of wide-dispersive emissions to air (TE5).

Emission to soil

The models for predicting soil emission include only 5 data points, i.e. chemicals. Still, the models have high squared correlation coefficients by calibration, but also high differences to the squared correlation coefficient by validation; the latter indicates lowered robustness. The F-ratio is high but influenced by the low number of observations. Best linear fit for local non-dispersive emissions (TE3) is by using production (D) as predictor parameter, while wide dispersive use (N) gives the best fit to the wide-dispersive soil emission (TE6).

The training set is used for cross-validating the emissions for each chemical by removing it from the training set and making a regression model with the remaining chemicals. Figure A2 shows the cross-validation in a scatter diagram between reported emissions and predicted emissions. Approximately half of the model predictions are overestimated relative to RA emissions and for TE1, TE2 and TE4 the overestimates are predominantly occurring for low emissions. For TE1 to TE5 the model predictions are within one order of magnitude (±) of the RA emissions for more than 75% of the chemicals. For TE6 63% of the chemicals are within one order of magnitude of the RA emissions. Conservative estimates are typically occurring for low reported emissions. The most reliable model in terms of number of cases (n) and Q^2 is wide-dispersive emissions to air, i.e. TE5.

References

1. Systat. Statistics. SPSS Inc., Chicago IL, 1997:751.
2. Mateu, J. Methods of Assessing and Achieving Normality Applied to Environmental Data (1997). Environmental Management, 21, 767-777.
3. Larsen, Pia Veldt. Regression and analysis of variance (2006) http://statmaster.sdu.dk/courses/st111/
4. CAMO ASA. The Unscrambler 9.02; Oslo, Norway, 2005.

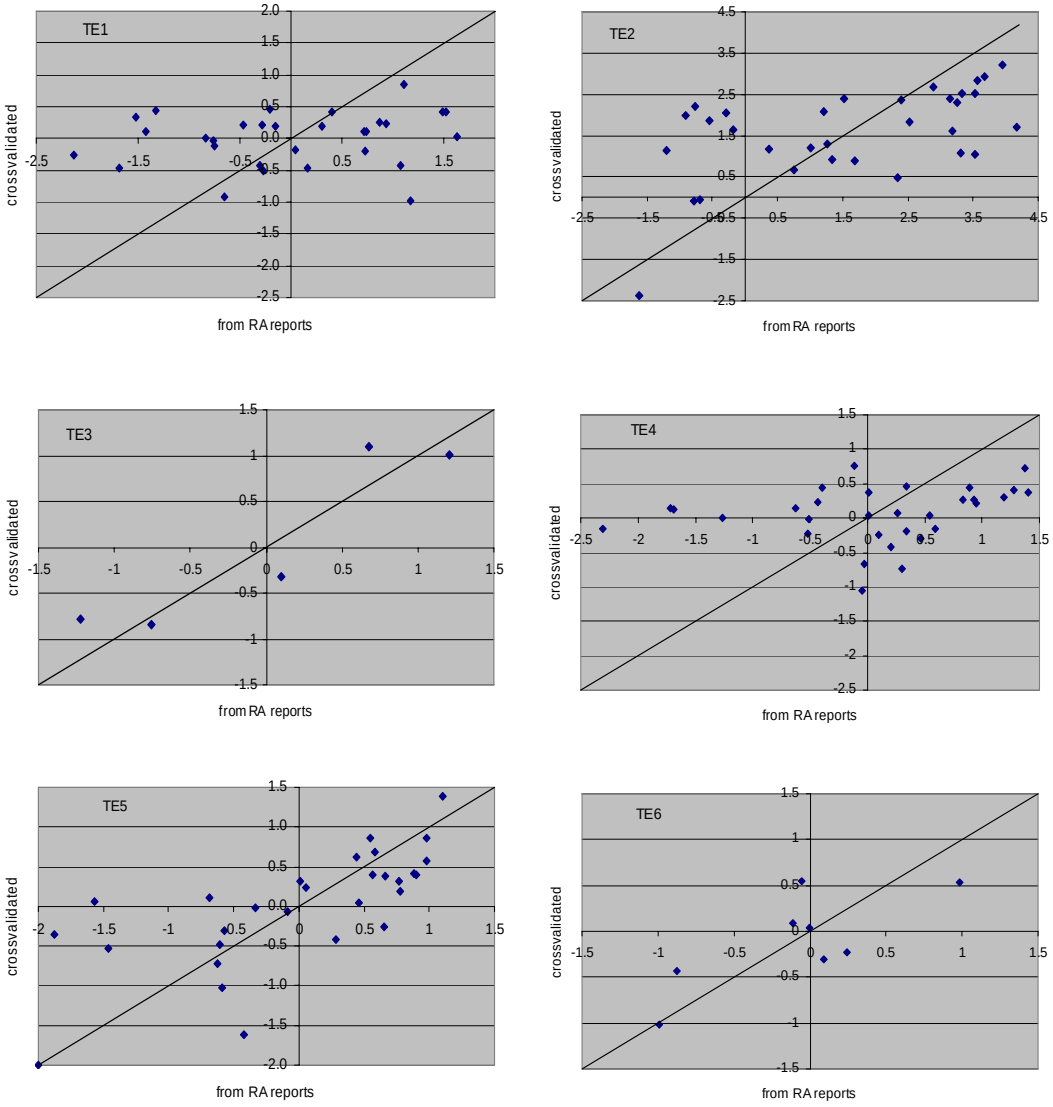


Figure A2 Scatter plots of reported emissions and leave-one-out cross-validated emissions for each of the six target emissions, TE1 to TE6. Model performance parameters are found in Table A1.

European Commission

EUR 24254 EN – Joint Research Centre – Institute for Environment and Sustainability
Title: Using decision tree analysis and GIS in Modelling (semi)VOC Emissions at the European Scale
Author(s): Fauser, P., Pistocchi, A., Thomsen, M.
Luxembourg: Publications Office of the European Union
2010 – 22 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-15023-4
doi:10.2788/62874

Abstract
Risk assessment of semi Volatile Organic Compounds (semi-VOCs) is a fundamental part in the regulation of production and use in industries and households. Emission inventories are a natural starting point in risk assessments and, given the complex use and emission patterns of the many thousands VOCs, emission estimates are often one of the most uncertain and problematic parts in risk assessments. Some critical issues are quantifying production and use amounts of chemicals and chemical containing products, assigning amounts to industrial activities and household products, identifying use and emission patterns, identifying receiving environmental compartment and quantifying the emission to these, which depend on production volumes, chemical properties, and their mode of use, in a non-trivial way.

To ensure reliable risk assessments, emission estimates are sought which need to be realistic and, at the same time, do not require excessive effort in the modelling of emission inventories.
The report proposes a method to capitalize on the information in the European Chemicals Bureau risk assessment reports (RARs), available for a limited number of chemicals, to train decision trees that allow estimating emissions of chemicals to different environmental compartments. The report also illustrates how these estimates can be used in conjunction with geographic information system (GIS) processing of spatial data to map emissions. Examples are drawn with reference to the case of the European Union. It is shown how quick, spatially distributed estimates of emissions to specific environmental compartments can be obtained to be used in screening level assessment.
The method outlined in the report allows a quick and reliable estimation of the fraction of total chemical production that results in emission to a specific environmental medium, using data mining techniques and GIS. This can result helpful within the new procedures for risk assessments guided by REACH, as a way to exploit data from existing risk assessments for predicting and mapping emissions of chemicals that have not yet been assessed.

Keywords: VOC, emissions, decision tree analysis, GIS, ECB reports

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

LB-NA-24254-EN-C